INVESTIGATION OF K-MEANS AND FUZZY K-MEANS CLUSTERING FOR THE ANALYSIS OF MASS SPECTROMETRY IMAGING DATA

A Thesis
Presented to
The Academic Faculty

by

Sanaiya Sarkari

In Partial Fulfillment
of the Requirements for the Degree
Biomedical Engineering in the
School of Engineering

Georgia Institute of Technology
December 2014

INVESTIGATION OF K-MEANS AND FUZZY K-MEANS CLUSTERING FOR THE ANALYSIS OF MASS SPECTROMETRY IMAGING DATA

Approved by:

Dr. May D. Wang, Advisor
School of Biomedical Engineering
Georgia Institute of Technology

Dr. John H. Phan
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: _____

[To the students of the Georgia Institute of Technology]

ACKNOWLEDGEMENTS

I wish to thank Dr. May D. Wang, Dr. Facundo Fernandez, Chanchala Kaddi and Rachel Bennett for their guidance and support during the research.

This research has been supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, the National Science Foundation (GRFP to CK), and the Georgia Institute of Technology Undergraduate Research Opportunities Program (PURA travel award to SS). We thank Dr. M. Cameron Sullards, Dr. Yanfeng Chen and Dr. Alfred H. Merrill, Jr. for input and for sharing the MSI data used in this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

MSI	Mass spectrometry imaging
MALDI	Matrix-assisted laser desportion ionization
FKM	Fuzzy k-means
CH	Calinski-Harabsz index
PCA	Principal component analysis

SUMMARY

Mass spectrometry imaging (MSI) is an experimental technique used to measure molecular composition across the surface of a sample, such as a tissue slice. MSI can simultaneously measure hundreds to thousands of molecules, and link those molecular profiles with their spatial location in the sample. However, MSI datasets can be very large, and identifying potentially important biological patterns is a challenging problem. Many types of explorative data analysis have been applied to MSI datasets, and in particular, k-means clustering has recently gained attention for this application [1]. In this study, we examine the effects of different parameters on the performance of basic k-means and fuzzy k-means clustering in identifying biologically relevant patterns in MSI datasets.

CHAPTER 1

INTRODUCTION

Mass Spectrometry Imaging

Mass spectrometry imaging (MSI) is a technique used to measure the molecular composition of a sample across its surface [8]. Previously, researchers would require prior knowledge about which molecule was important and had to carefully stain tissues sections for specific markers [4]. MSI increased popularity in the biochemistry field is attributed to its ability in picking out and mapping spatial arrangement of thousands of molecular species at once

Despite much contribution to the scientific community, MSI still possess many challenges. Two main regions of focus for current research lies in reducing computational and statistical challenges in analyzing MALDI-MS images and establishing a pipeline that can use used to analyze unsupervised data [2]. In 'MALDI imaging mass spectrometry: statistical data analysis and current computational challenges', Alendandrov proposes steps to apply when approaching unsupervised MALDI-MSI data for analysis through clustering [3].

Clustering Algorithms

Clustering, spatial segmentation used to detect molecular expression patterns, has become a popular explorative method for analyzing MSI data. Clustering works by dividing samples into a selected number (k) of clusters based on the similarity of samples to cluster centers. 'k' points are randomly selected to represent initial centroids. Next each pixels or objects are grouped to the closest centroid. Once all pixels are assigned,

locations for ‘k’ centroids are recalculated. The previous two steps are repeated until no more convergence is able to take place.

Examples of clustering algorithms include basic k-means clustering, and fuzzy k-means (FKM) clustering. Both basic k-means and fuzzy k-means are used to form compact cluster. Basic k-means is sensitive to outliers and noise and only use numerical attributes. Fuzzy k-means is a general iterative clustering method that is adaptations form the basic k-means algorithm that serves to minimize intra-cluster variance.

Incorporation of clustering algorithm and MSI

Still during its infancy stage, spatial segmentation of most MALDI – MSI data, especially with oncology application, was performed with the use of hierarchical clustering. ‘Spatial and Spectral Correlation in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis’ by McCombie et al. concludes that clustering with the aid of principal component analysis (PCA) and linear discriminant analysis (DA) helps identify spatial correlation of mass spectra [5]. However, it is noted that each clustering yields only moderately for specific clusters. Article form Deininger et al. also found that when comparing MALDI-MSI (that have undergone PCA with hierarchical clustering) to histology of cancerous specimen, not always are results completely congruent [4].

In ‘MALDI-imaging segmentation is a powerful tool for spatial functional proteomic analysis for human larynx carcinoma’, Alexandrov uses basic k-means for its faster processing on larynx carcinoma data to differentiate tumorous and non-tumorous regions [1]. Similar experimentation was reproduce in his later paper titled ‘MALDI imaging mass spectrometry: statistical data analysis and current computational

challenges' with addition of component analysis, and increased cluster size on unsupervised rat brain data set [3]. Alexandrov's latter research found that the method suggested does not always provide clear regions of interest and as compared to the component analysis only one spatial pattern is shown despite the fact that a peak can contribute several spatial patterns [3]. In order to solve this, Alexandrov performs fuzzy k-means in addition to basic k-means only to find that results from fuzzy k-means are similar to that of crisp clustering [3].

There are two distinct aims of this paper. While previous literature concentrates on finding an approach for extracting meaningful images for unsupervised dataset through a specific clustering algorithm, we aim to aid in the knowledge by comparing the effects of both fuzzy k-means clustering to basic k-means clustering and see of the two which is best suited to extract informative patterns in MALDI – MSI. Subsequently, we intend on looking at the effects that different distance metrics such as square Euclidean, city block, correlation and cosine have on the basic k-means algorithm. Quantitative comparison for both aims will be carried out using Calinksi Harabasz index (CH index) and correlation between MS images.

CHAPTER 2

PROCEDURE

Sample Preparation

We implement clustering algorithms on three different MALDI MSI datasets of mice brain tissue. The first dataset describes the cerebellum of a mouse model of Tay Sachs/Sandhoff disease (spatial dimensions: 91x84, spectral dimension: 4,438) [Chen et al., 2008]. The second and third datasets contains coronal (spatial dimensions: 103x169, spectral dimension: 8,000) and sagittal (spatial dimensions: 104x168, spectral dimension: 8,000) views of a healthy mouse brain, respectively [Bennett et al., 2013]. All necessary algorithms and evaluation metrics are performed in MATLAB.

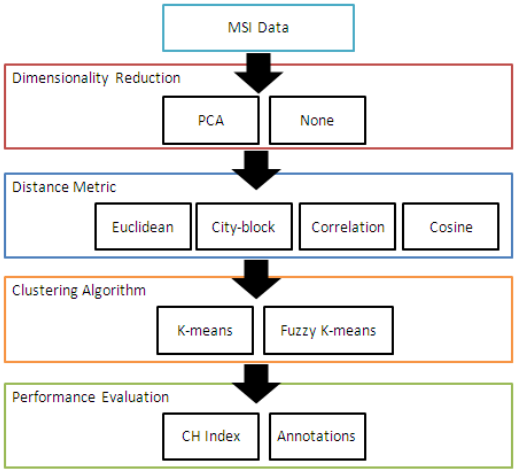

Figure 1. Flow chart showing workflow: the effects of dimensionality reduction, distance metrics, and clustering algorithm were examined.

Clustering: K-means and Fuzzy K-means

In order to perform a systematic comparison between clustering algorithms to identify relevant patterns in the MSI dataset, two types of clustering algorithms have been looked into: basic k-means and fuzzy k-means. The k-means algorithm works by partitioning dataset into ‘k’ clusters. Each data point is assigned to the cluster closest to it

which is calculated with a specified distance metric. The previous steps are repeated until all data points are fixated to a specific cluster. While there are many distance measurement for basic k-means, this paper looks at square Euclidean, city block, correlation and cosine. The square Euclidean works similar to that of the Euclidean distance metric, but it does not take the square root. City block, also known as Manhattan distance metric, tracks distance between data points by following a grid-like path. The correlation distance metric measures similarities between profiles. Lastly, cosine measures similarity between two vectors by measuring the cosine of angle between them. Basic K-means is a commonly used technique since it is typically computationally faster and produces tighter clusters than hierarchical clustering. Despite its advantage, it is very difficult to measure the quality of clusters produced since different initial partitions direct affect the outcome.

Fuzzy k-means is a variation of k-means clustering in which each sample in a data set is assigned to each for the ‘k’ clusters with a certain probability. With Fuzzy K-means only square Euclidean distance metric was used.

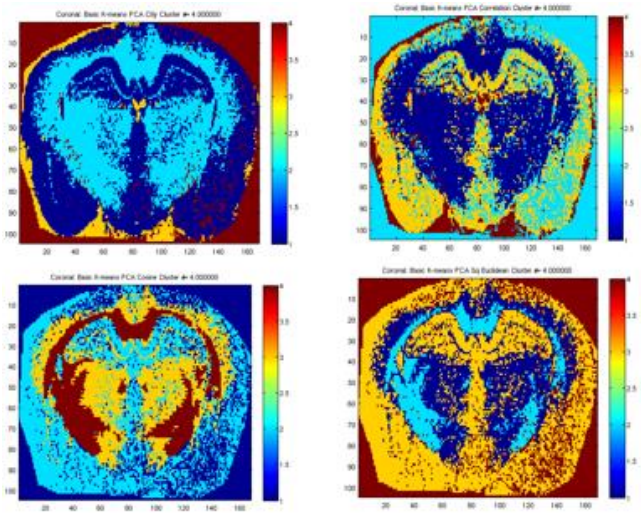

Figure 2. K-means clustering results (k = 4) using the coronal dataset. Top: city-block (l), correlation (r), Bottom: cosine (l), and Euclidean (r).

Component Analysis

Component analyses are heavily used in MSI data set to extract meaningful tissue images while having no prior understanding of the underlying histology. This paper specifically looks into the effects of including principal component analysis. PCA is typically applied to identify patterns in data and express the similarities and difference between the high dimensional dataset where good graphical representation is not available. PCA works by initially subtracting the mean from each data set dimension, which is the average across each dimension. Next, the covariance matrix is calculated, followed by calculation of eigenvectors and eigenvalues of the covariance matrix. It is these eigenvectors of the covariance matrix that characterize the data. Finally, chose components and form a feature vector. Doing so should indicate that eigenvector with the highest eigenvalue is the principle component of the dataset. Once a feature vector is formed and transposed a new data set is derived. PCA is a highly regarded choice since once it finds patterns; you can compress the data and reduce the number of dimensions, without losing much information, increasing computational speed to analyze the MALDI – MS images. Comparisons will be made to all datasets undergoing both clustering algorithms with and without PCA.

Evaluation Metrics

The first evaluation metric used in this paper is the Calinski Harabasz index (CH index), an intrinsic evaluation that measure cluster quality based on how the clusters are. The index is defined as the sum between clusters over sum within cluster multiplied by $\frac{N-k}{k-1}$, where N is the number of observations and k represents the number of clusters. Subsequently, the second form of evaluation metric will be correlation. A total of 21 m/z

images were handpick for correlation comparison and can be found in supplementary data. Once the MS images have undergone clustering with or without PCA, each image a specific cluster (all from 2 to 10) then is correlated to each of the 21 images, pixel by pixel. Once an array of correlation values are extracted the mean values are looked into for a comprehensive comparison

CHAPTER 3

RESULTS

In order to compare the effects that the different distance metrics would have on the clustering algorithm, all four distance metrics are compared against each other for both data sets with or without component analysis.

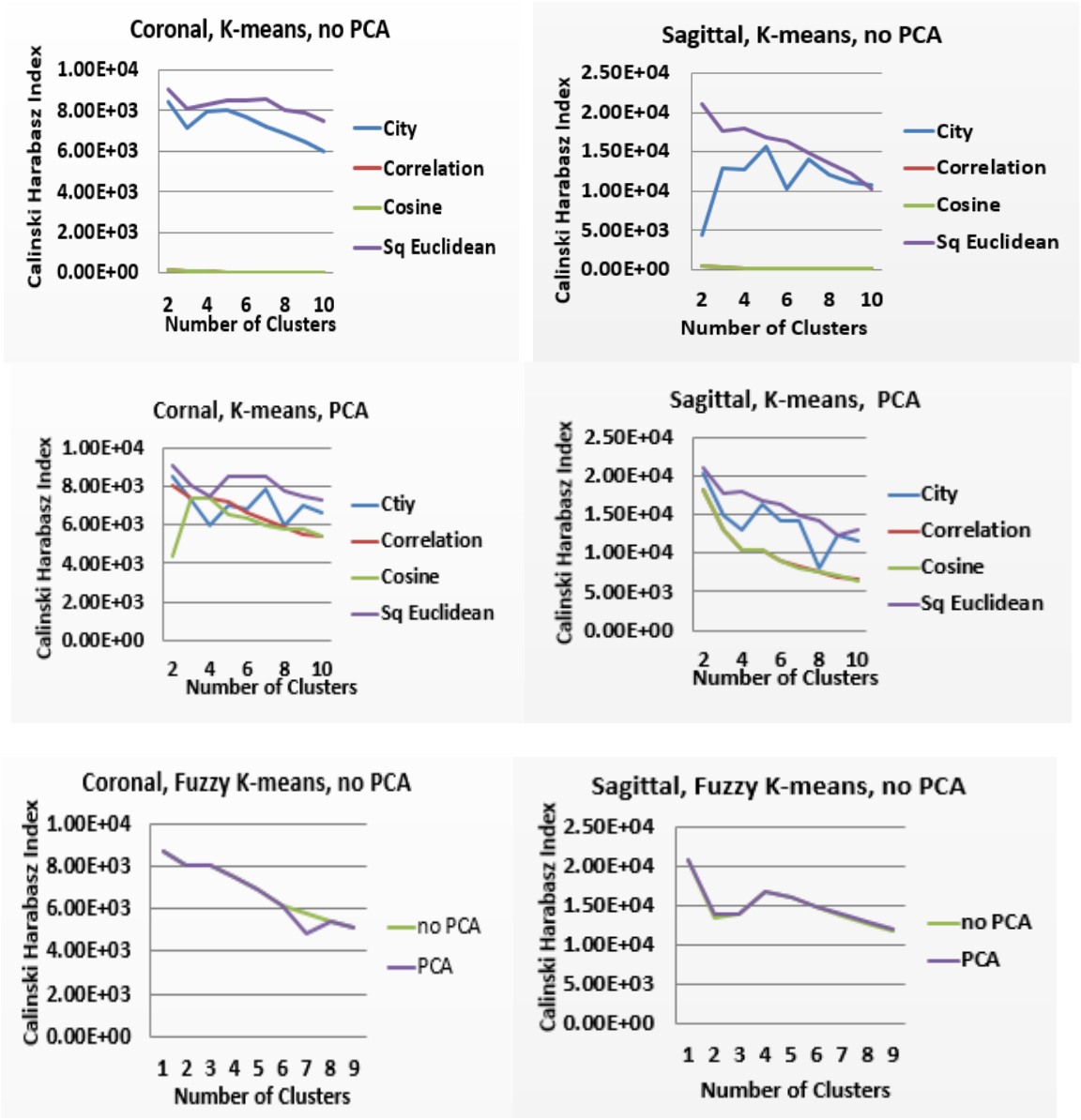

Figure 3. Calinksi-Harabasz index analysis: Variations in the Calinski-Harabasz index across different clustering pipelines in the coronal and sagittal datasets.

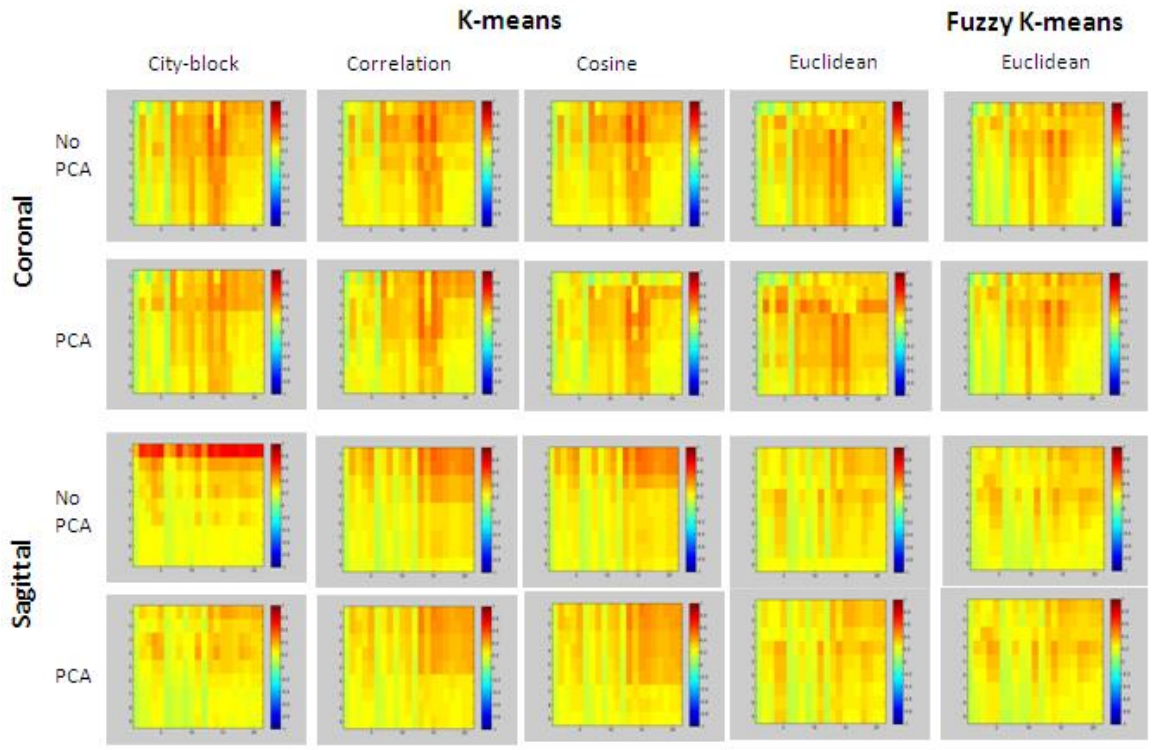

Figure 4. Mean correlation analysis: Mean correlation of the binary cluster images with the 21 m/z value images selected via manual annotation.

Coronal, Basic K-means, without PCA

All distance metrics yield optimal CH index values at K =2 with CH index values of 8400 (city), 135(correlation), 136(cosine), 9050(square Euclidean). Holistically, square Euclidean distance metric contained the highest CH index value in comparison to the others. At K = 7, square Euclidean also contains a peak that contains a CH index of 8560 which is close to its optimal CH index value. Both cosine and correlation distance metrics maintain similar magnitudes for the CH index. Correlation maintains CH index between 135.64 and decreases to 20.91, whereas cosine starts at 135.62 and decrease to 19.80. City block ranges from 8400 to 5990.

Coronal, Basic K-Means, with PCA

City block, square Euclidean and correlation contain optimal CH index for cluster size of 2 with CH index values of 8560, 8060, and 9050 respectively. Cosine maintains the highest CH index at K = 4. Ranges for CH index for distance metric include city (5950 to 8560), correlation (5430 to 8060), cosine (4340 to 7430) and square Euclidean (7330 to 9050).

Sagittal, Basic K-Means, without PCA

Correlation, cosine and square Euclidean have optimal CH index at cluster size of 2, whereas city has the highest CH index at cluster size of 5 and the lowest at 2. Ranges included: city (4350 to 15600), correlation (81.79 to 377.20), cosine (80.45 to 379.83) and square Euclidean (10200 to 21000).

Sagittal, Basic K-Means, with PCA

All distance metrics yield optimal CH index values at K =2 with CH index values: city (20400), correlation (18300), cosine (18300) and square Euclidean (21000). While correlation, cosine and square Euclidean have the lowest CH index at K = 10, city contains the lowest CH index at K = 8. Ranges included: city (4350 to 15600), correlation (81.79 to 377.20), cosine (80.45 to 379.83) and square Euclidean (10200 to 21000).

Coronal, Fuzzy K-Means, with and without PCA

Contains optimal clustering at K = 2: 8740 for both with and without PCA. While both contain lowest CH index at similar magnitudes at K = 10, applying PCA does have a higher CH index of 5100, whereas without PCA contains CH index of 5090.

Sagittal, Fuzzy K-Means, with PCA

Similar to the results with coronal FKM, application of PCA resulted in minute changes. Highest CH index are maintained at K = 2 for with application of PCA (21000) and without PCA (21000). At K = 10, clustering both PCA and without PCA yields lowest CH values of 11800 and 12000 respectively.

In an effort to see which cluster would perform the closest fit, images of clustering were compared to the original m/z image through correlation. Doing so, it is evident that for the coronal data set, correlation distance metric without PCA holistically had higher correlation for throughout all clusters. Likewise, square Euclidean had a similar effect with PCA.

For sagittal data, city block distance metric, with PCA indicates that closest clusters take place at K = 2, and K = 6 without PCA. For correlation and cosine distance metric, both cases without PCA holistically shows a higher correlation each with tight clusters at 8 and 2 respectively with the addition of PCA. Square Euclidean displays a strong correlation for cluster 10 without PCA and 2 with PCA.

Coronal dataset for fuzzy k-means did not have significant correlation, but sagittal with fuzzy k-means yields strong correlation at K = 8 for without presence of PCA and K = 10 with presence of PCA.

CHAPTER 4

DISCUSSION

After comparing CH index for both basic k-means and fuzzy k-means it is evident that basic k-means produces images that contain the tightest clusters as seen by the high CH index values. This can be especially seen in instance where K = 2 consist of optimal clustering since the algorithm is differentiating between white and gray matter. Basic K-means with square Euclidean distance metric yields the tightest clusters as seen from the CH index values, closely followed by city block distance metric.

In comparison to the effect PCA had in the datasets between basic k-means and fuzzy k-means, it is evident that the presence of PCA significantly played an impact when applied in conjunction to basic k-means with cosine and correlation distance metric as seen in both datasets. Additionally, PCA drastically altered the optimal CH index for city block distance metric at K = 2 for one dataset. Despite its contribution towards k-means dataset, addition of PCA had little to no impact when applied with fuzzy k-means.

K = 2 was found to contain the tightest cluster as well as have the highest correlation to the original m/z images amongst several combinations. As mentioned earlier this can be attributed to the distinction between gray and white matter having more differentiation as oppose to clusters within each specific matter.

CHAPTER 5

CONCLUSION

Higher number of clusters reveals distinct spatial patterns within the sample. Square Euclidean consistently presented the highest Calinski Harabasz index. Future works will concentrate on comparisons to additional data sets including DESI-MSI while investigating other distance metrics and extensions to k-means clustering such as Hamming and Harmonics k-means, respectively.

REFERENCE

[1] K. Schwamborn and R. M. Caprioli, "Molecular imaging by mass spectrometry - looking beyond classical histology," *Nature Reviews Cancer,* vol. 10, pp. 639-646, Sep 2010.

[2] T. Alexandrov, "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges," *Bmc Bioinformatics,* vol. 13, Nov 2012.

[3] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis," *Analytical Chemistry,* vol. 80, pp. 9649-9658, Dec 2008.

[4] R. M. Parry, A. S. Galhena, C. M. Gamage, R. V. Bennett, M. D. Wang, and F. M. Fernandez, "OmniSpect: An Open MATLAB-Based Tool for Visualization and Analysis of Matrix-Assisted Laser Desorption/Ionization and Desorption Electrospray Ionization Mass Spectrometry Images," *Journal of the American Society for Mass Spectrometry,* vol. 24, pp. 646-649, Apr 2013.

[5] T. Alexandrov, M. Becker, O. Guntinas-Lichius, G. Ernst, and F. von Eggeling, "MALDI-imaging segmentation is a powerful tool for spatial functional proteomic analysis of human larynx carcinoma," *Journal of Cancer Research and Clinical Oncology,* vol. 139, pp. 85-95, Jan 2013.

[6] T. Alexandrov, M. Becker, S. O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, and P. Maass, "Spatial Segmentation of Imaging Mass Spectrometry Data with Edge-Preserving Image Denoising and Clustering," *Journal of Proteome Research,* vol. 9, pp. 6535-6546, Dec 2010.

[7] T. Alexandrov and J. H. Kobarg, "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering," *Bioinformatics,* vol. 27, pp. I230-I238, Jul 2011.

[8] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis," *Analytical Chemistry,* vol. 77, pp. 6118-6124, Oct 2005.

[9] S. O. Deininger, M. P. Ebert, A. Futterer, M. Gerhard, and C. Rocken, "MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers," *Journal of Proteome Research,* vol. 7, pp. 5230-5236, Dec 2008.

[10] B. Desgraupes, "Clustering Indices". Retrieved from http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

[11] Sarkari, S., Kaddi, C., Bennett, R., Fernandez, F., & Fernandez, F. (2014). Comparison of Clustering Pipelines for the Analysis of Mass Spectrometry Imaging Data. *IEEE*.